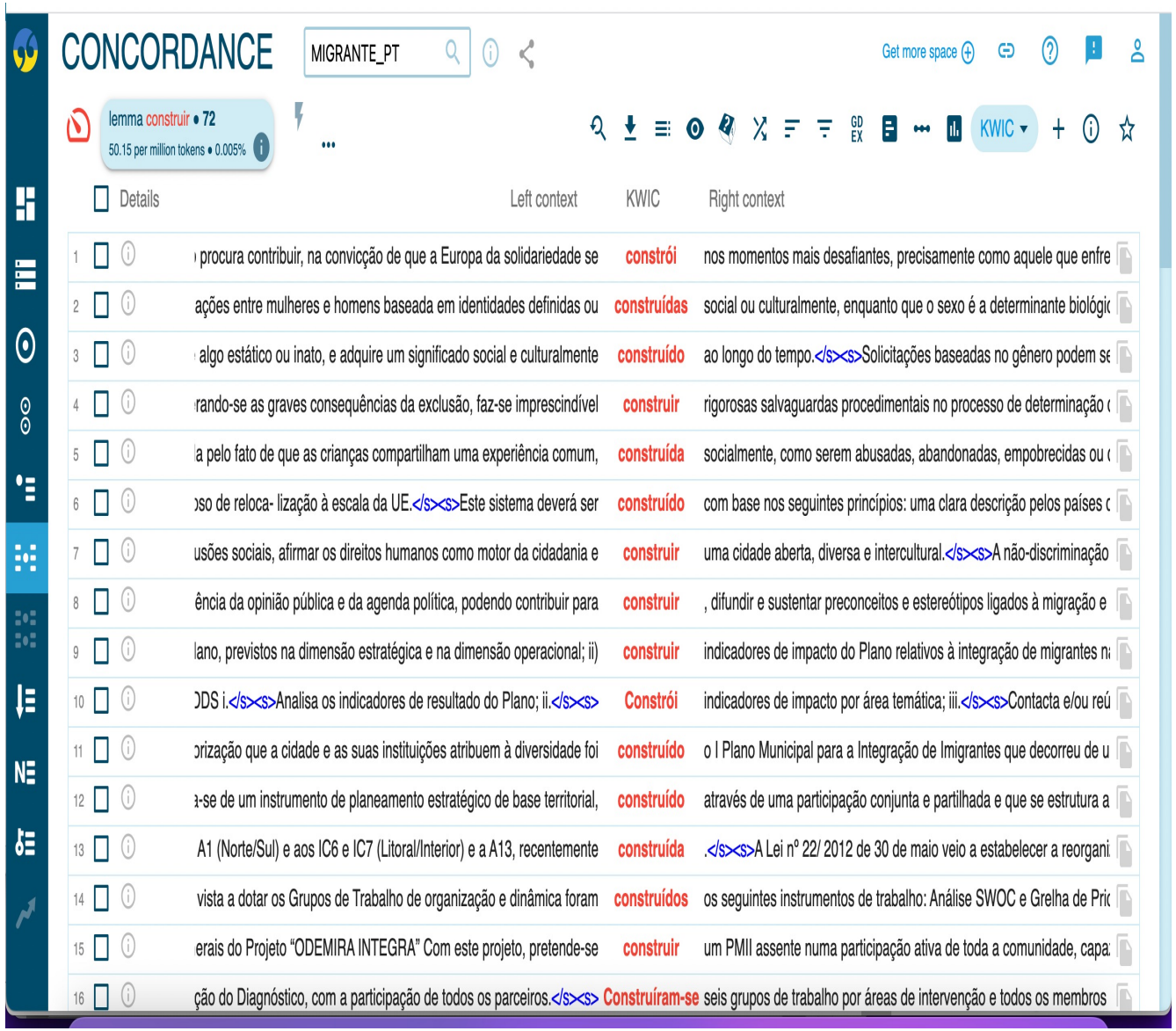
Constituição de *corpus*: critérios, parâmetros e representatividade



CENTRO DE LINGUÍSTICA DA UNIVERSIDADE NOVA DE LISBOA

Constituição de corpus

CLUNL
CENTRO DE LINGUÍSTICA DA UNIVERSIDADE NOVA DE LISBOA

Esquema de trabalho

1. Constituição do *corpus*
    - seleção
    - amostragem
    - normalização
    - codificação
    - descrição do *corpus*

Constituição de corpus

CLUNL
CENTRO DE LINGUÍSTICA DA UNIVERSIDADE NOVA DE LISBOA

Criação do *corpus*: questões iniciais (Sinclair, 1996)

- Concepção do *corpus*: linguistas *vs.* especialistas?
  - depende dos objetivos

- Formatos: oportunidades e dificuldades
  - formato digital inicial
  - digitalizações (OCR e interpretação)
  - transcrição (discurso oral e interpretação)

Constituição de corpus

Criação do *corpus*: questões iniciais (cont.)

- Permissões e direitos de autor: meios e estratégias

  - depende da utilização: investigação *vs*. divulgação

    → contratos de licenças (autores e editoras)

    → obras livres de direitos de autor (70-80 anos)

    (não necessariamente edições livres de direitos de autor)

    → autorizações dos entrevistados ou responsáveis

Constituição de corpus

Constituição do *corpus* e seleção

- A seleção do que entra para um *corpus* depende dos objetivos de utilização do *corpus*.

  - discurso escrito e discurso oral?
    - → *corpus* de referência
    - → *corpus* literário
    - → *corpus* de especialidade

  - como e onde obter discurso oral?
  - discurso formal e informal? Como e onde obter...

Constituição de corpus

- A seleção do que entra para um *corpus* depende dos objetivos de utilização do *corpus*. (cont.)

  - tipologia de textos e amostragem?
    - artigos revistas, teses, notícias? Quantos de cada?

  - textos típicos/representativos?
    - de quê? representatividade estatística *vs*. representatividade de outros elementos

    → tipicalidade: padrões linguísticos e resultados normativos

(→ tipicalidade e critérios de amostragem)

Constituição de corpus

- A seleção do que entra para um *corpus* depende dos objetivos de utilização do *corpus*. (cont.)

  - período de tempo considerado

  - tamanho (do *corpus*, das subcoleções, das amostras...)

  - amostragem vs. textos integrais (tipos de *corpora*)

  - critérios mínimos, informação e codificação
  (tipo de texto, autor, sexo, idade, local...)
  → redesenho e 'especialização' do *corpus*

Constituição de corpus

Critérios de seleção (gerais)

- **Tipos de texto**

Corpus geral/de referência

Corpus literário: subgénero (romance, novela, poesia...)
Corpus de especialidade: ...?

- artigos
- comunicação técnica interna
- materiais pedagógicos
- materiais divulgação (especialista – especialista; especialista – não-especialista)

Constituição de corpus

Critérios de seleção (gerais)

- **Período**
  - 20 anos?
  - 50 anos?
  - ..
- **Língua**
  - original?
  - tradução?
  - língua materna do autor?

Constituição de corpus

Critérios de seleção (gerais)

- **Tamanho**
  - das amostras
  - do corpus

"The dimension of a corpus are of prime concern for most researchers in the initial conceptualization, and in the public statements. In the long run, they matter very little. The only guidance I would give is that a corpus should be as large as possible and should keep on growing." (Sinclair 1991: 18)

Sinclair, J. (1991). "Corpus, Concordance, Collocation", Oxford University Press.

Constituição de corpus

Critérios de seleção (outros)

- **Publicação/edição**
  - local
  - empresa?
  - …
- **Grau de escolaridade(?)**
  - licenciatura
  - doutoramento
  - título de especialista?

Constituição de corpus

CLUNL
CENTRO DE LINGUÍSTICA DA UNIVERSIDADE NOVA DE LISBOA

Métodos de seleção

- Critérios de seleção normativa (canonicidade/tipicidade/representatividade estatística)
- Critérios de seleção metódica (amostragem da população)
- Critérios de seleção orientados para os resultados (contexto da investigação/constituição do *corpus*)

Constituição de corpus

Métodos de seleção (cont.)

- **Critérios de seleção normativa**
  (canonicidade/tipicidade/representatividade estatística)

Canonicidade: estatuto de prestígio social, cultural, económico que reflete legitimação consensual e normativa.

→ Noção temporal e culturalmente marcada!

Constituição de corpus

Métodos de seleção (cont.)

- **Critérios de seleção metódica** (amostragem da população)

- amostragem estatística (externa e não interna...)

externa (composição do *corpus*)

quantos e quais os textos a selecionar de ente uma subcoleção?

→ amostragem aleatória vs.

amostragem ***representativa*** **de diversidade**?

Constituição de corpus

CLUNL
CENTRO DE LINGUÍSTICA DA UNIVERSIDADE NOVA DE LISBOA

Métodos de seleção (cont.)

- **Critérios de seleção orientados para os resultados** (contexto da investigação/constituição do *corpus*)

- critérios mistos/abordagem pragmática:
    - amostragem
    - tipicidade/canonicidade
    - disponibilidade (formato, direitos de autor, …
    - objetivos de investigação (ex.: diversidade…)

Constituição de corpus

CLUNL
CENTRO DE LINGUÍSTICA DA UNIVERSIDADE NOVA DE LISBOA

Constituição de *corpus* e amostragem

- "elegibilidade" (nível 1: critérios de seleção)

- "composição" (nível 2: critérios/intervalos das amostras do corpus/por subcoleção...)

1. de 10% a 50% de textos escritos por doutorados
2. 10 autores representados por 3 textos
3. pelo menos 30% de textos relevantes (ex. revistas indexadas) e pelo menos 30% de textos não relevantes
4. pelo menos 20% de textos curtos (1500 a 6 000 palavras) e pelo menos 20% de textos longos (mais de 80 000 palavras)

...

→ ***equilíbrio do corpus***

Constituição de corpus

Normalização e codificação

‘Data is ontologically different from the world.’

Moisl (2009: 876)

Ex.: um texto digitalizado e em formato eletrónico é diferente do texto original em papel

- formatação (tipo de letra, itálicos, maiúsculas, negritos…)
- translineação (divisão de palavras por linha)
- quebras de página
- …

Constituição de corpus

Normalização e codificação (cont.)

- A normalização implica decisões a aplicar sobre TODOS os textos sobre

  - formatos
  - alterações
  - codificação de elementos no original
  - codificação de alterações feitas…

→ a normalização pode ser mais ou menos relevante

dados linguísticos?

Constituição de corpus

CLUNL
CENTRO DE LINGUÍSTICA DA UNIVERSIDADE NOVA DE LISBOA

Normalização e codificação (cont.)

- Exemplo:
  - mantemos/retiramos o número da página?
  - corrigimos erros ortográficos? *voçê*? e gralhas? *exmplo*? (indicamos essa correção?)
  - mantemos/retiramos os hífens da translineação?
  - mantemos/alteramos a grafia original?

→ *implicações a nível da estatística lexical...*

Constituição de corpus

Normalização e codificação (cont.)

- Exemplo:

  - mantemos/identificamos o itálico/negrito/diferentes tipos de letra?

    → *implicações a nível da identificação de neologismos, estrangeirismos, ênfase…*

  - mantemos/normalizamos aspas diferentes?

    → *implicações a nível da identificação de neologismos, estrangeirismos, ênfase, intertextualidade…*

Constituição de corpus

CLUNL
CENTRO DE LINGUÍSTICA DA UNIVERSIDADE NOVA DE LISBOA

Codificação (vs. anotação)

- A **codificação** diz respeito à expressão e marcação de informações inerentes aos textos

  - codificação de elementos ao nível supra textual (cabeçalho do texto; cabeçalho do corpus)

  - codificação de elementos ao nível do texto

  (parágrafos, formatos, cortes, etc...)

- A **anotação** diz respeito à expressão e marcação de informações extra texto (anotação morfossintática, etc...)

Constituição de corpus

CLUNL
CENTRO DE LINGUÍSTICA DA UNIVERSIDADE NOVA DE LISBOA

Codificação

- CABEÇALHO DO TEXTO *metadados (exemplo CRPC)*

Natureza dos dados
Sigla da natureza dos dados
Fonte
Sigla da fonte
Número de ordem
Nome do autor
Ano de nascimento do autor
Sigla do autor
Nome do jornal/revista
Sigla do nome do jornal/revista
Título
Número do volume
Sigla do título
Nome da disciplina curricular
Sigla da disciplina curricular
Ano de escolaridade
Secção
Número da edição
Número do jornal/revista
Editor
Sigla do editor
Colecção
Sigla da colecção
Localidade da edição
Sigla da localidade
Data
Sigla da data
Género/Tema
Sigla do género/tema

Constituição de corpus

CLUNL
CENTRO DE LINGUÍSTICA DA UNIVERSIDADE NOVA DE LISBOA

Codificação

- CABEÇALHO DO TEXTO *metadados (exemplo CRPC)* (cont.)

Página
Coluna
País da edição
Sigla do país
Directoria
Ficheiro
Número de linhas
Número de palavras
Número de caracteres
Estado
Observações
Introdução
Correcção
Revisão
Data da 1ª edição

País do autor
Língua materna do autor
País de nascimento do autor
Local de nascimento do autor
Endereço web do download
Data do download
Língua do original
Situação de Direitos de Autor

É possível codificar informação no nome de ficheiro e nome das pastas!

Constituição de corpus

Descrição do *corpus* e codificação

- Elemento essencial para a utilidade e relevância do *corpus*
- Discriminação dos elementos constituintes do corpus, de acordo com os critérios de seleção e amostragem definidos.

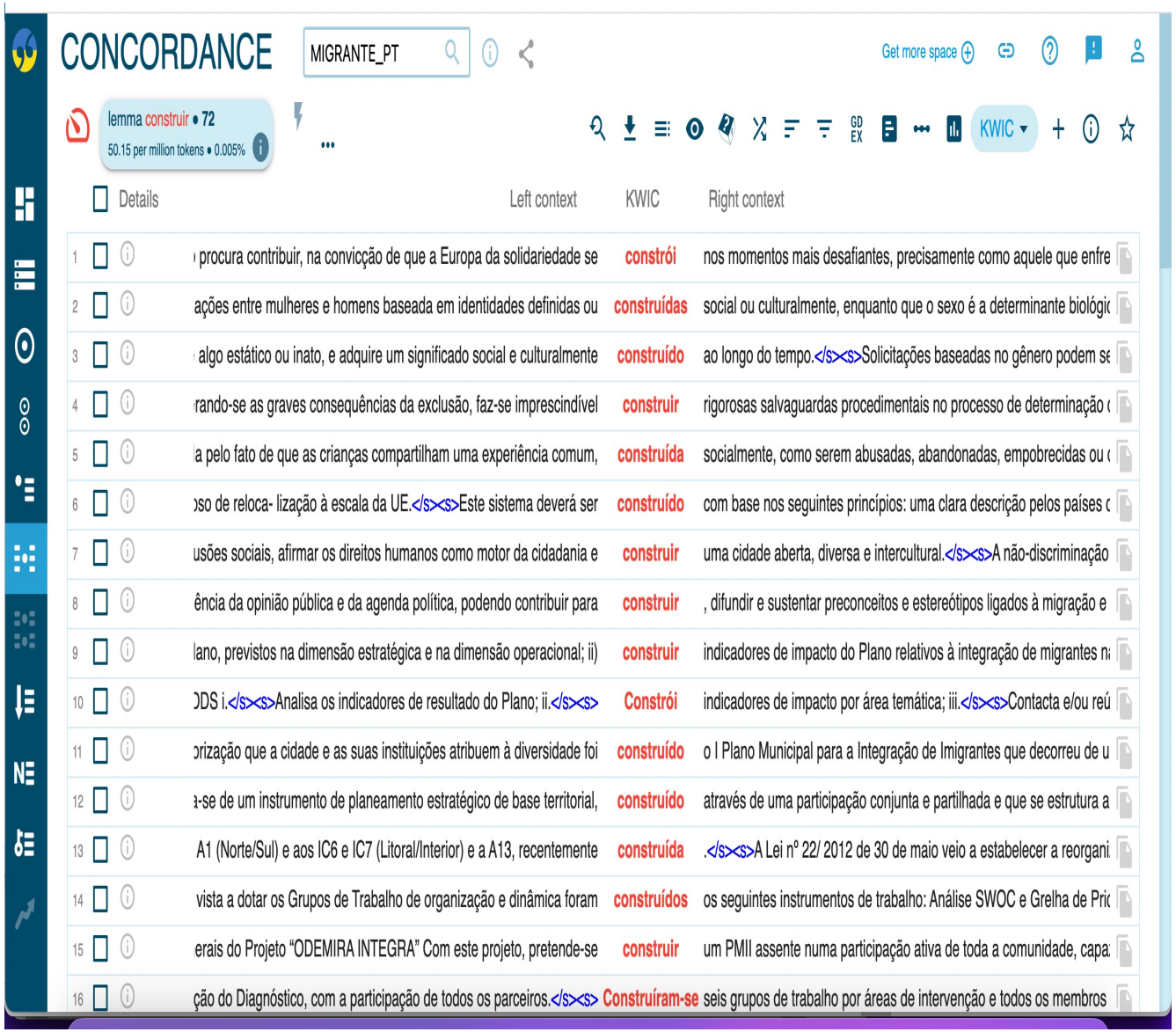
> "Until we know a lot more about the effects of our design strategies, we must rely on publishing a list of exactly what is in a corpus (...). Users and critics can then **consider the constitution and balance of the corpus as a separate matter from the reporting of the linguistic evidence of the corpus**."
>
> (Sinclair 1991:13)

Constituição de corpus

- Critérios de elegibilidade
- Critérios de amostragem
- Normalização
- Codificação e descrição do corpus

→ **permitem a avaliação do corpus!**

Obrigada!



CENTRO DE LINGUÍSTICA DA UNIVERSIDADE NOVA DE LISBOA

Erasmus+

NOVAFCSH

EN	The European Commission's support for the production of this publication does not constitute an endorsement of the contents, which reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.
PT	O apoio da Comissão Europeia à produção desta publicação não constitui um aval do seu conteúdo, que reflete unicamente o ponto de vista dos autores, e a Comissão não pode ser considerada responsável por eventuais utilizações que possam ser feitas com as informações nela contidas.